Forecasting the 2024 U.S. Presidential Election through Poll Aggregation and Adjustments for Poll Quality*

A Hybrid Model Predicts a Narrow Lead for Kamala Harris in Swing States as Election Day Approaches

Colin Sihan Yang Lexun Yu Siddharth Gowda

November 4, 2024

This paper develops a hybrid model to forecast the 2024 U.S. presidential election by combining poll aggregation techniques with additional variables, including pollster reliability, sample size, timing relative to the election, geographic region, and polling methodology. By integrating these factors, our model accounts for variations in poll quality and regional differences in voter sentiment, providing a more stable and accurate prediction compared to traditional poll aggregation alone. Our results show that Kamala Harris holds a slight lead over Donald Trump in most swing states, with support levels stabilizing closer to election day. This approach underscores the importance of nuanced forecasting models in election analysis, revealing how poll quality and timing can significantly influence support predictions. Ultimately, this model offers a comprehensive tool for understanding the dynamics of voter sentiment and improving the accuracy of election predictions.

1 Introduction

Election result forecasting has become an essential tool for analysts in political science and the public to predict the outcome of democratic process, such as the presidential election in the United States. Traditionally, individual polls have been used as a snapshot of voter sentiment, but they only reflect temporary changes in the performance of contestants, instead of a precise estimation of the election result. As discussed by Pasek (2015) and Blumenthal (2014), the aggregation of multiple polls, or "poll-of-polls," has become a popular technique to

*Code and data are available at: https://github.com/yulexun/uselection.

reduce individual survey errors and provide more accurate election forecasts. However, conventional aggregation models often overlook dynamic election factors, including the credibility and quality variations among poll sources.

We build a hybrid election forecasting model following the strategies mentioned by Pasek (2015). As Pasek (2015) described in their article, aggregation involves determining which surveys are worth including, as well as selecting, combining and averaging results from multiple polls to reduce individual biases and errors. Our model incorporates this approach, filtering polls from FiveThirtyEight (2024) based on a numeric grade that indicates each pollster's reliability.

The objective of this study is to predict the support percentages for the primary candidates, Kamala Harris and Donald Trump, using this hybrid model. We incorporate aggregation by filtering the polls on FiveThirtyEight (2024) by numeric grade that indicates pollster's reliability. We also quantify other polling attributes, such as sample size, poll reliability scores, and polling methodology, which serve as predictors. Our results show that this model provides a more stable and accurate forecast than traditional poll aggregation alone.

The remainder of this paper is structured as follows: Section 2 provides an overview and exploration of the data. Section 3 provides the modeling approach, including simple and multiple linear regression models and a Bayesian hierarchical model. We then present our results in Section 4 and discuss the implications, limitations, and future research directions in Section 5.

The data gathering and analysis is done in R (R Core Team 2023) with the following packages: knitr (Xie 2014), tidyverse (Wickham et al. 2019), ggplot2 (Wickham 2016), dplyr (Wickham et al. 2023), arrow (Richardson et al. 2024), here (Müller 2020), gridExtra (Auguie 2017), Matrix (Bates, Maechler, and Jagan 2024), Rstan (Stan Development Team 2024) and lubridate (Grolemund and Wickham 2011).

1.1 Estimand

The estimand for this research paper is the predicted support percentages for Kamala Harris and Donald Trump. The prediction is based on quantifying various polling factors, including sample size, poll scores, and transparency scores, which are used as predictors.

2 Data

2.1 Measurement

The dataset we obtained from FiveThirtyEight (2024) is accumulated from multiple polls and surveys. According to FiveThirtyEight, they aggregate polling data conducted by other firms

and organizations that meets their methodological and ethical standards (Morris 2024). The dataset contains a list of questions in polls and surveys and their results. When a new poll is conducted, the poll is appended to the dataset. FiveThirtyEight assign the poll a poll_id, and each question is assigned a question_id.

In each of the polls recorded by FiveThirtyEight, all options of each question are recorded in candidate_names, while the proportion of respondents choosing that option is recorded in pct as a percentage. The pct is our primary response variable.

There are limitations in these measurements. The differences in sampling methods, wording in survey questions and systematic biases are all reflected in the outcome (Radcliffe and Morris 2023). Thus, in the dataset, FiveThirtyEight includes other variables, such as pollscore and sample_size for their prediction model in addition to the polling results for transparency and accuracy. As an example, pollscore indicates the pollster's reliability rating, a low pollscore indicates a higher reliability rating (Silver 2008).

In this dataset, each row represents a polling question that records the variables of interest. Each entry allows us to explore the real-world relationships between polling factors and the support percentage (`pct`) for the candidates Kamala Harris and Donald Trump. This dataset enables an analysis of how various polling characteristics influence the reported support levels for the candidates we are focused.

These variables combined allow researchers to reliably analyze and predict the 2024 US election result over time.

2.2 Data Exploration

The raw data from FiveThirtyEight contains 52 columns, all of the column headers are displayed below:

poll_id	pollster_id	pollster
sponsor_ids	sponsors	display_name
pollster_rating_id	pollster_rating_name	numeric_grade
pollscore	methodology	transparency_score
state	start_date	end_date
sponsor_candidate_id	sponsor_candidate	sponsor_candidate_party
endorsed_candidate_id	endorsed_candidate_name	endorsed_candidate_party
question_id	sample_size	population
subpopulation	population_full	tracking
created_at	notes	url
url_article	url_topline	url_crosstab
source	internal	partisan
race_id	cycle	office_type
seat_number	seat_name	election_date
stage	nationwide_batch	ranked_choice_reallocated
ranked_choice_round	hypothetical	party
answer	candidate_id	candidate_name
pct	poll_id	pollster_id

These columns can be categorized into three types, response variable, numeric predictors and categorical predictors.

The response variable is:

- pct: The percentage of support or vote share that each candidate (Kamala Harris or Donald Trump) received in the poll.

Example of numeric predictors are:

- sample_size: The total number of respondents in each poll.
- pollscore: A numeric score representing the error and bias of the pollster. Negative numbers are better.
- numeric_grade: A numeric grade assigned to each pollster, reflecting pollster quality or reliability.

Example of categorical predictors are:

- state: The U.S. state in which the poll was conducted or targeted.
- methodology: The method used to conduct the poll (e.g., online, phone, in-person).
- party: The political party of the candidate (e.g., DEM for Democrat, REP for Republican).
- candidate_name: The name of the candidate, either Kamala Harris or Donald Trump.
- pollster: The name of the polling organization that conducted the poll.
- stage: The stage of the election (e.g., “general”).

2.3 Clean Data

The data cleaning process involves several steps to ensure the quality and relevance of the polling data. First, we filter the dataset to retain only poll results with a numeric grade of 2.7 or higher, indicating that the polls are considered reliable. Next, we address missing values in the state attribute: polls with NA in the state column are considered national polls.

We then create a new attribute, days_taken_from_election, which represents the time gap between the poll's start date and the actual U.S. election date. Additionally, we filter the dataset to include only polls conducted after July 21, 2024, the date when Kamala Harris declared her candidacy. Finally, we remove any remaining rows that contain missing values to ensure a clean dataset.

2.3.1 States included in anaylsis

After the data cleaning process, 21 states had no polling data. A table showing the number of polls for each state, including those without any polls, is provided in Table 4.

This absence of polling data is not a significant concern due to the structure of the United States Electoral College (explained in detail in the Appendix). The states lacking polling data have consistently followed historical voting patterns, so predicting the winning candidate in those states is unnecessary.

2.4 Cleaned data

The first 6 rows of the dataset is displayed in Table 1.

Table 1: Sample of cleaned US election data

pct	Sample Size	pollscore	Days taken from election	state	methodology	Candidate name
47.6	4180	-0.8	24	National	Online Ad	Kamala Harris
50.7	4180	-0.8	24	National	Online Ad	Donald Trump
0.8	4180	-0.8	24	National	Online Ad	Jill Stein
0.1	4180	-0.8	24	National	Online Ad	Chase Oliver
0.1	4180	-0.8	24	National	Online Ad	Cornel West
48.1	4180	-0.8	24	National	Online Ad	Kamala Harris

2.4.1 Basic Statistics Summary for Cleaned Data

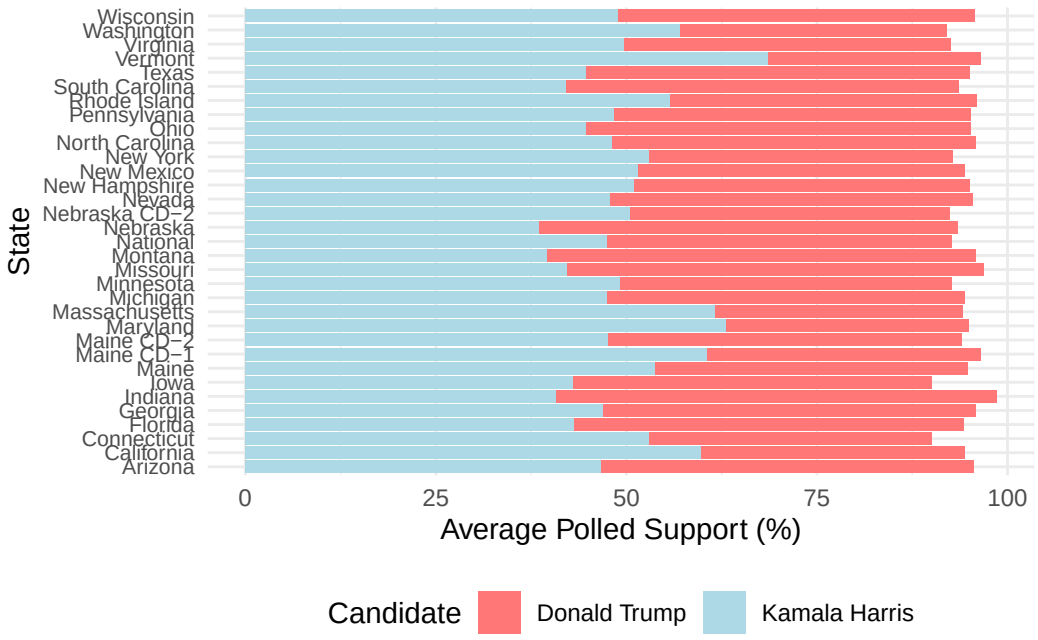

Figure 1: Historical State Voting Trends Are Maintined in 2024

In Figure 1, historically Democratic are polling for Kamala and historically Republican states are polling for Trump. Similarly, historically swing states also appear to be close, for instance Michigan (46.9% Trump, 47.5% Harris), Nevada (47.7% Trump, 47.9% Harris), and Pennsylvania (46.9% Trump, 48.2% Harris) all are close to an even split.

Figure 2 shows that polls utilizing multiple communication methods to reach voters tend to have lower interquartile ranges (IQRs) in their boxplots compared to those that rely on only one or two communication methods.

2.4.2 Relationship Between Variables

Figure 3 illustrates a weak positive correlation between a pollster's pollscore and the sample size of their poll. The figure also shows that most polls have a sample size around 800 to 1200 participants. It is also important to note that a few polls with exceptionally large sample sizes were excluded from the graph due to their status as clear outliers.

Based on Figure 4, there does not seem to be a relationship between the sample size of a poll the percentage of a candidate.

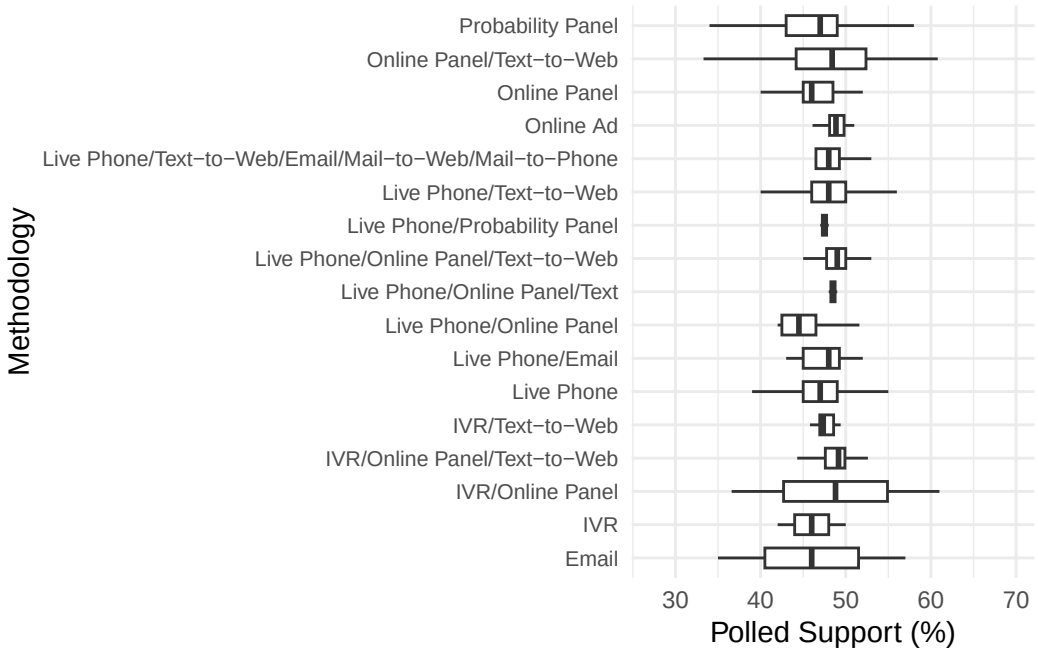

Figure 2: Polls with more Methodologies have Less Variability.

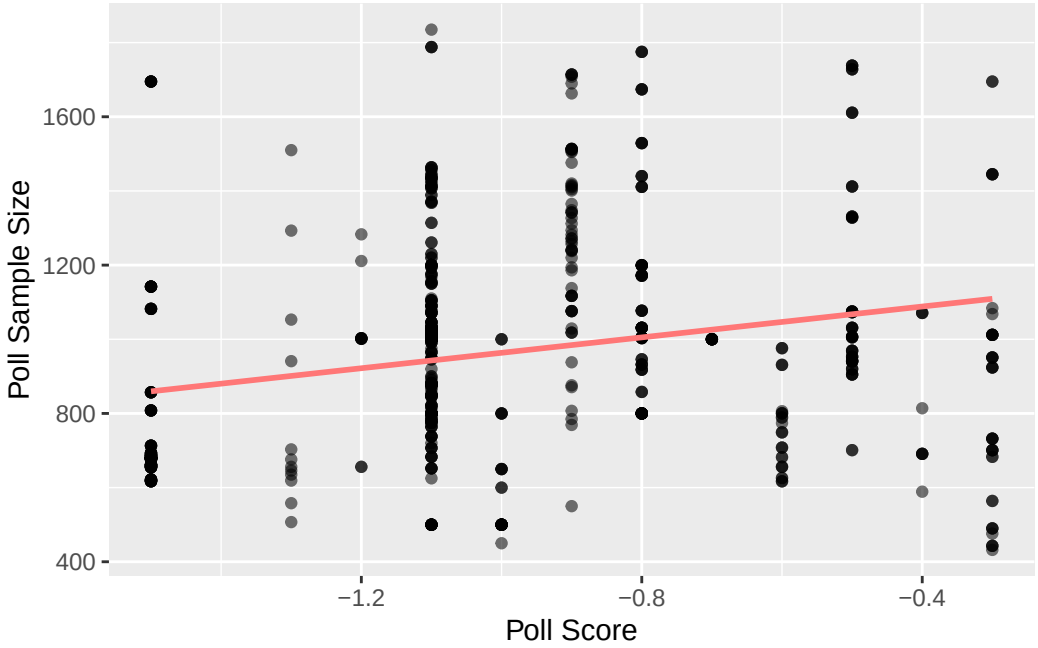

Figure 3: More Reliable Pollsters Have Larger Sample Sizes in their Polls

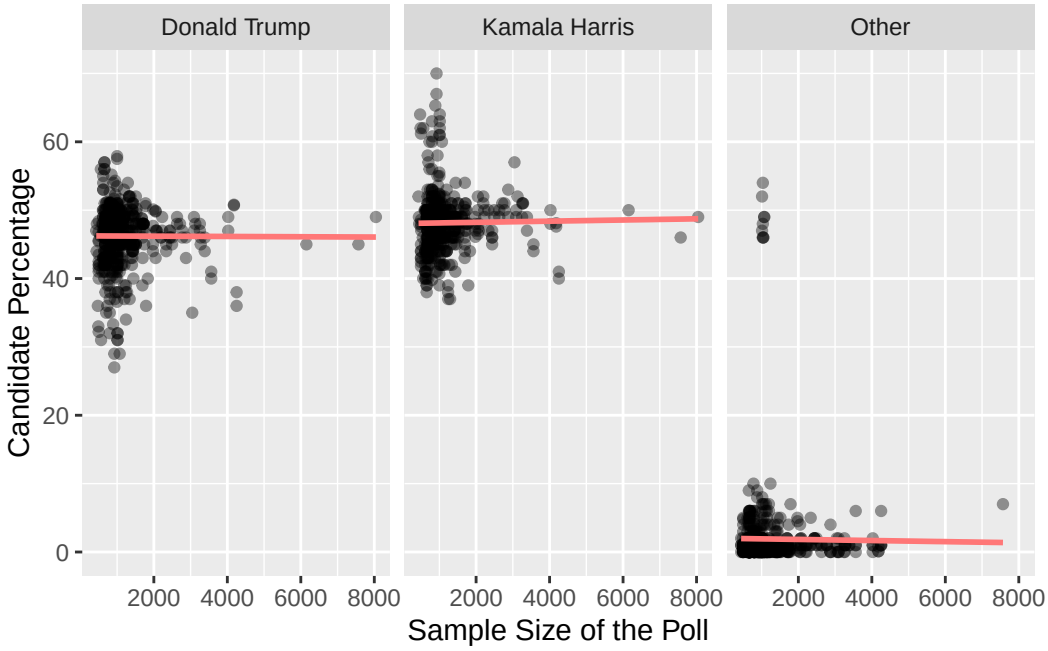

Figure 4: The Sample Size of a Poll does not impact a Candidate's Voting Percentage

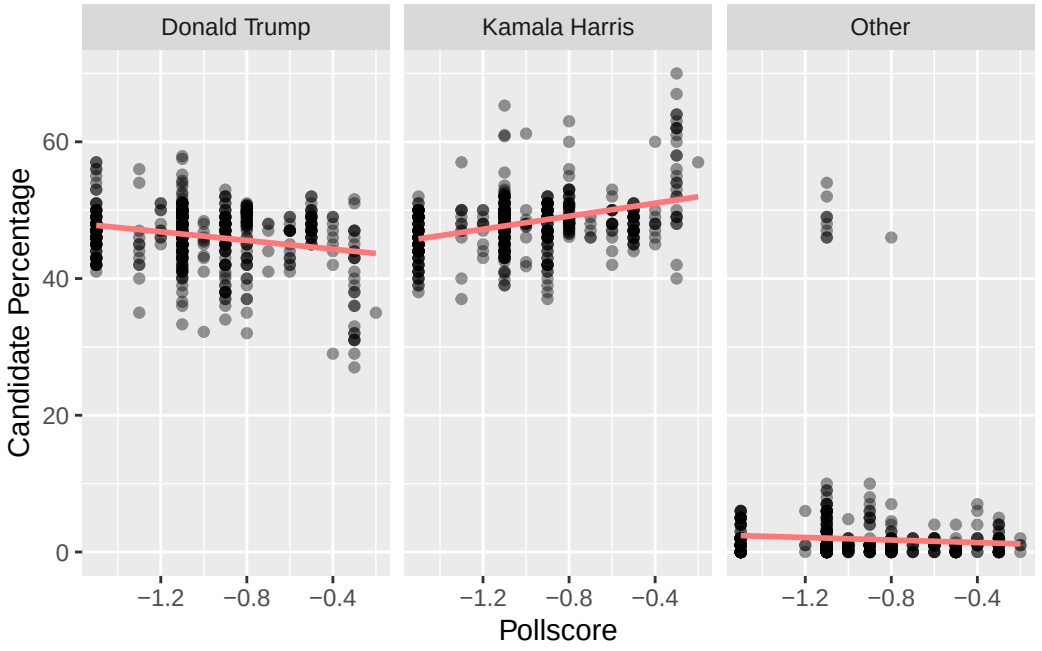

Figure 5: More Reliable Pollsters Score Higher For Trump

Figure 5 depicts the relationship between a pollster's poll score and a candidate's percentage. For Donald Trump, a negative correlation is observed, indicating that pollsters with lower poll scores tend to assign him a higher percentage of support. Conversely, Kamala Harris shows the opposite trend: as the poll score decreases, her percentage tends to rise. This suggests that more reliable polls, characterized by lower poll scores, report higher support for Trump compared to less reliable pollsters.

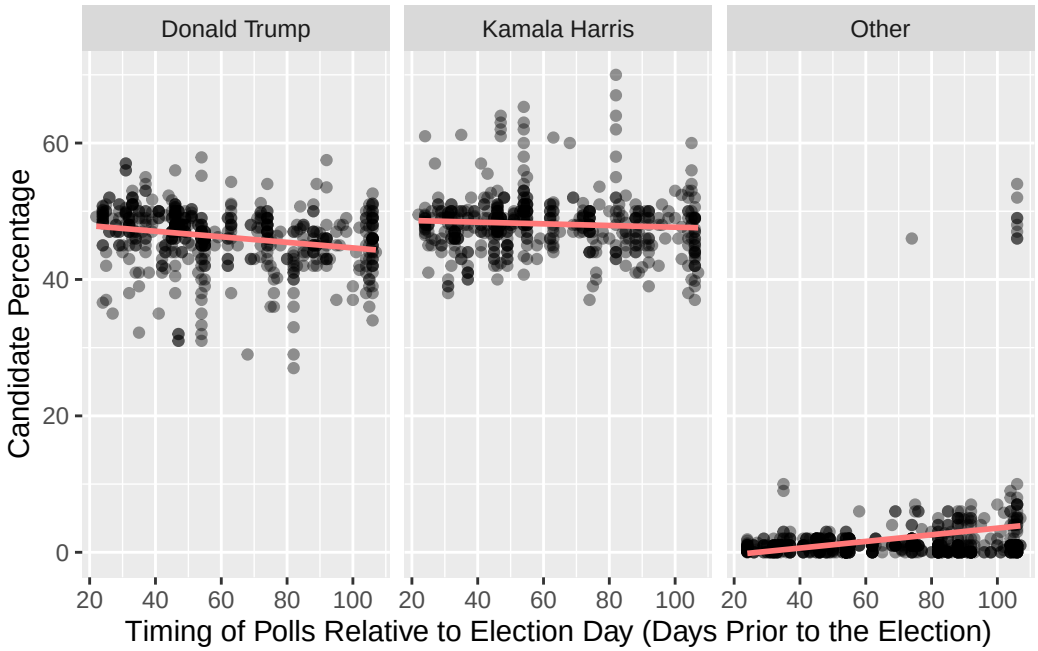

Figure 6: Both Canidate Percentages Are Gaining Support Closer to the Election

Based on Figure 6, Trump and Harris are getting more votes in polls that are done closer to the election. This is a result of non-major candidate support rapidly decreasing. Specifically, Trump support is increasing at a faster rate than Harris. However, the polls in general show a slight lead for Harris throughout the last 100 days.

The rapid decline in third party support could be due to Robert F. Kennedy dropping out of the race (this sentences should probably be in results or discussion section).

3 Model

The goal of our modeling strategy is twofold. Firstly, we aim to accurately predict the support percentage (PCT) for Harris and Trump based on relevant poll data and key influencing factors. Secondly, we seek to evaluate the efficacy of different modeling approaches—from simple linear regression (SLR) to multiple linear regression (MLR) and Bayesian hierarchical

models—to understand their predictive capabilities and assess the underlying relationships between variables. By comparing these models, we can determine which approach provides the most robust and reliable predictions, while considering the variability and potential uncertainty in the data.

3.1 Model set-up

The Bayesian model is implemented in R (R Core Team 2023) using the rstanarm package as described by Brilleman et al. (2018). The model is run with the following specifications:

- Formula: pct $\sim$ pollscore + days taken from election + sample size + (1|methodology) + (1|state)
- Priors: Normal(0, 2.5) for all coefficients and intercept, Exponential(1) for σ
- Settings: Seed = 123, Cores = 4, Adapt delta = 0.95

We run the model in R (R Core Team 2023) using the `rstanarm` package of Brilleman et al. (2018). We use the default priors from `rstanarm`.

3.2 Basic SLR Model

We first set up a SLR model for predicting pct based on pollscore. The model is:

$\hat{\text{pct}} = \beta_0 + \beta_1 \times \text{pollscore}$

In this SLR model, the response variable is pct and the only one predictor is the pollscore. Figure 7 visualizes the relationship between the actual and predicted values of percentage support (pct) for Kamala Harris, based on a simple linear regression model with pollscore as the sole predictor. Blue points represent individual comparisons between actual and predicted values. The red dashed line represents the line of perfect prediction, where actual values would equal predicted values.

The primary concern lies in the evident dispersion of data points, which are widely spread and do not cluster closely around the line of perfect prediction (the dashed red line). This suggests that while pollscore may have some predictive capability, it does not adequately explain the variability in pct. The observed inconsistencies between actual and predicted values indicate that the relationship between pct and pollscore is likely not sufficiently captured by a linear model with just one predictor.

In Figure 8, the distribution of the points suggests that the model might not capture much variability in pct when sample_size is the only predictor. This indicate that sample_size has limited impact on pct.

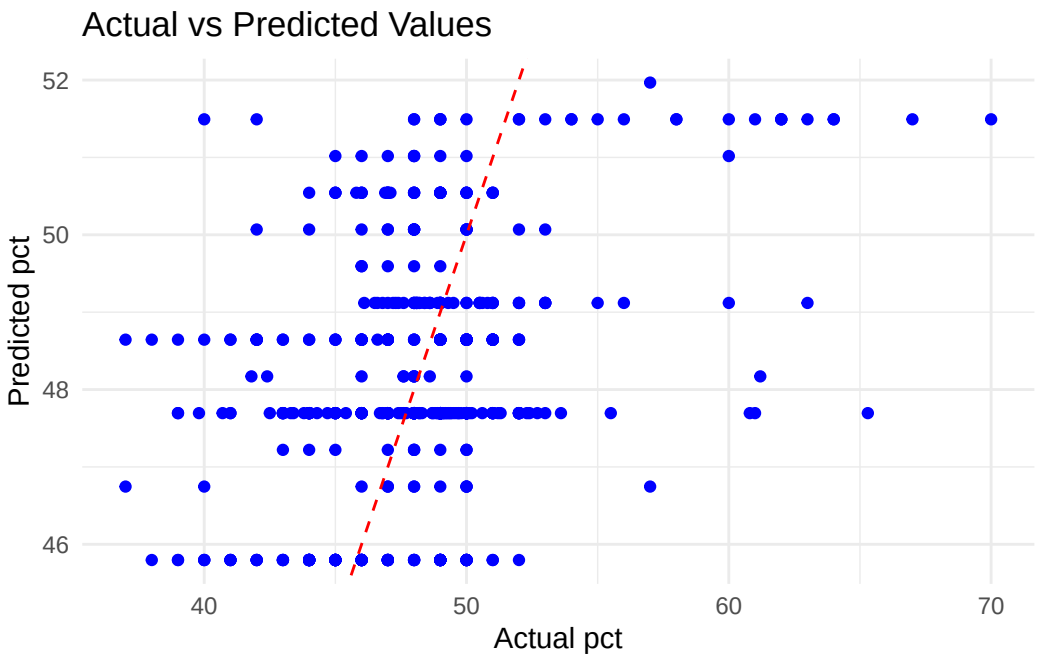

Figure 7: Comparison of Actual vs Predicted Percentage Support for Kamala Harris (Single Predictor: Pollscore)

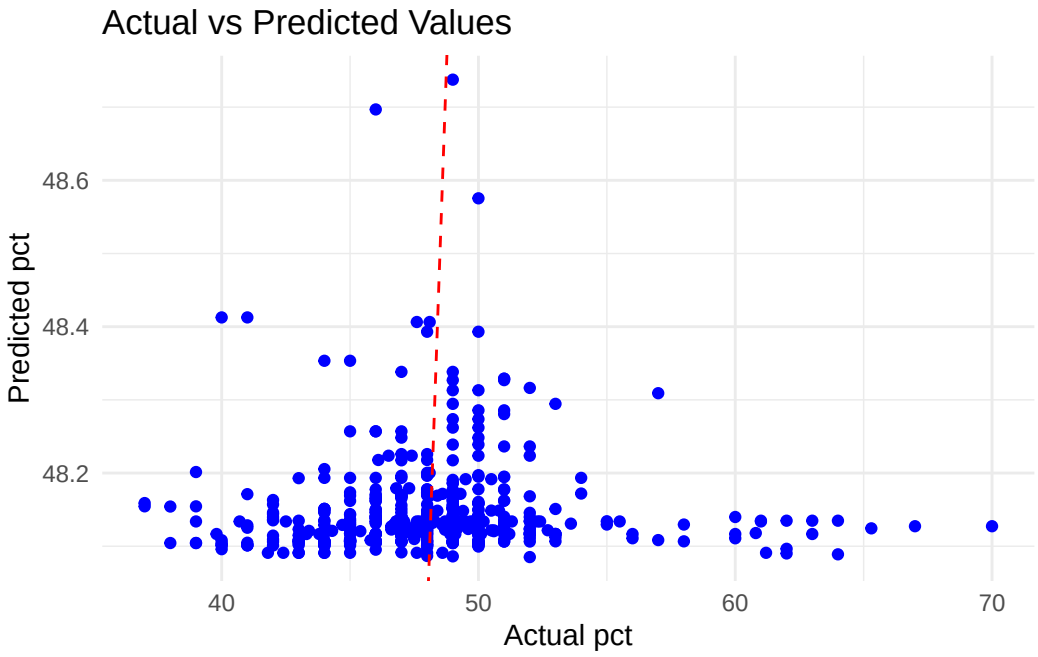

Figure 8: Actual vs. Predicted Values for pct Using sample_size as Predictor

3.3 MLR model

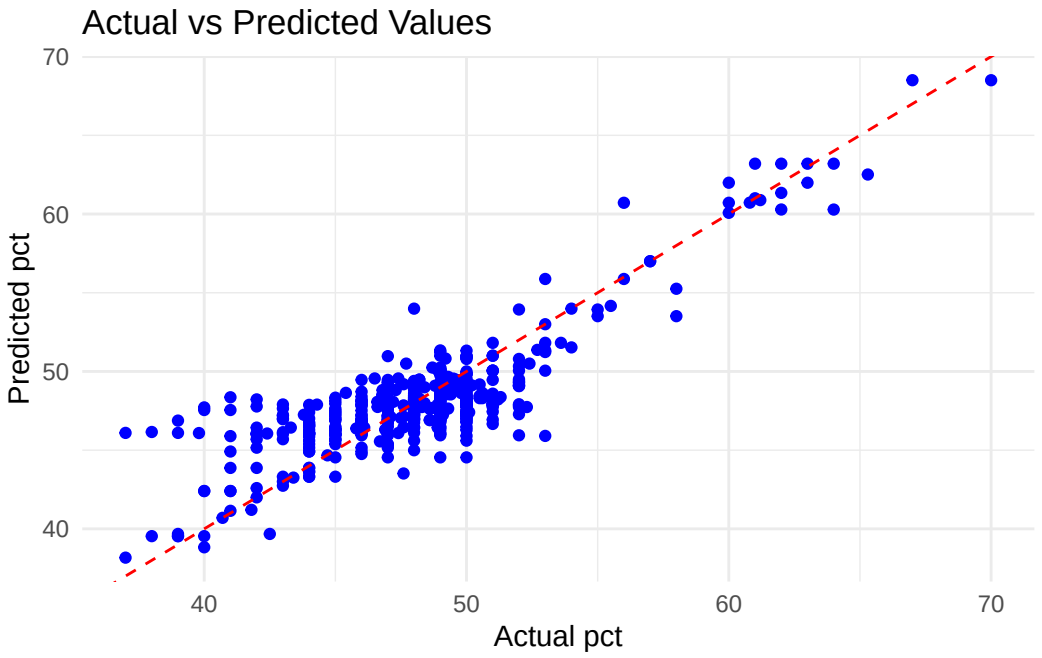

Figure 9: Actual vs Predicted Percentage Support for Kamala Harris (Multiple Predictors)

Next, we predicts pct using pollscore, days teken from election, methodology, sample size, and state as predictors. The model is:

$$\hat{\text{pct}} = \beta_0 + \beta_1 \cdot \text{pollscore} + \beta_2 \cdot \text{days\_taken\_from\_election} + \beta_3 \cdot \text{methodology} + \beta_4 \cdot \text{sample\_size} + \beta_5 \cdot \text{state}$$

Figure 9 compares actual pct values with the predicted pct values from the MLR model. Blue points represent individual comparisons between actual and predicted values. The red dashed line represents the line of perfect prediction, where actual values would equal predicted values.

The overall distribution of points suggests that the MLR model has captured a substantial portion of the variance in pct. The majority of the data points align relatively well with the red dashed line, particularly within the middle range of pct values (approximately between 40 and 60). This alignment indicates that the model performs reasonably well in this range, with predicted values correlating strongly with the actual observed outcomes.

3.4 Bayesian Model

Next, we set up a bayesian model as the following in accordance to Section 3.1:

$$
\begin{aligned}
\text{pct}_i &\sim \mathcal{N}(\mu_i, \sigma) \\
\mu_i &= \alpha + \beta_1 \cdot \text{pollscore}_i + \beta_2 \cdot \text{days\_taken\_from\_election}_i + \beta_3 \cdot \text{sample\_size}_i \\
&+ u_{\text{methodology}[i]} + u_{\text{state}[i]} \\
\alpha &\sim \text{Normal}(0, 2.5) \\
\beta_j &\sim \text{Normal}(0, 2.5), \quad j = 1, 2, 3 \\
u_{\text{methodology}} &\sim \text{Normal}(0, \sigma_{\text{methodology}}) \\
u_{\text{state}} &\sim \text{Normal}(0, \sigma_{\text{state}})
\end{aligned}
$$

This hierarchical Bayesian model is designed to capture both the fixed effects of predictors and the random effects of grouping factors. The response variable is pct (percentage of support), predictors are pollscore, days taken from election, and sample size. A normal prior with mean 0 and standard deviation 2.5 is set for all coefficients and the intercept, scaled automatically.

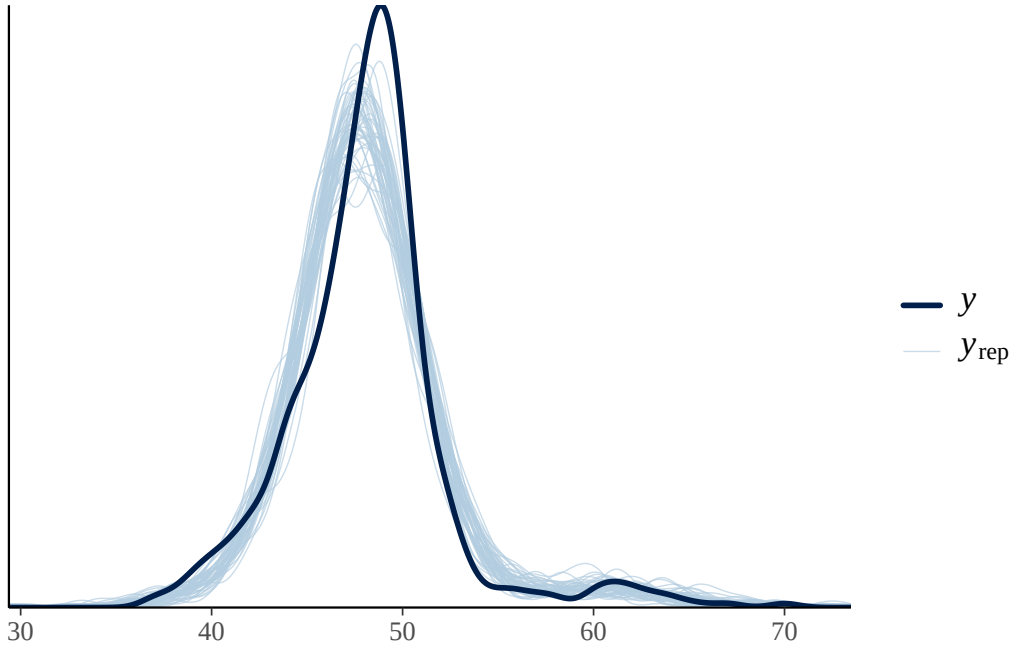

Figure 10: Posterior Predictive Check for the Bayesian Model Predicting pct

3.4.1 Posterior Predictive Checks

Figure 10 displays the results of a posterior predictive check (PPC) for a Bayesian model using the pp_check() function.

The observed data (y) follows a distinct, symmetrical distribution centered around a specific value, suggesting a well-defined peak with tapering on both sides. The replicated distributions (y_rep), shown as multiple thin lines, generally align with the shape of the observed distribution, indicating that the Bayesian model has captured the main characteristics of the data. However, the variability in the y_rep lines highlights the degree of uncertainty inherent in the model's predictive capability.

The fact that the y_rep curves closely match the overall pattern of the actual y suggests that the model performs reasonably well in replicating the observed data. Minor discrepancies or deviations between the y and y_rep might imply areas where the model could be fine-tuned or adjusted to improve accuracy. Overall, this PPC indicates that the model provides a decent fit to the data, with some variability accounted for in the predictive samples.

3.4.2 Train Test Validation

we implemented a train-test split validation approach to evaluate the performance of a Bayesian hierarchical model predicting percentage support (pct) for Kamala Harris. The dataset was divided into training (70%) and test (30%) sets to fit and validate the model, respectively. We used the stan_glmer function to fit the model on the training data, with predictors including pollscore, days_taken_from_election, sample_size, and random intercepts for methodology and state. After training, we generated predictions on the test set and calculated an R squared value to quantify the model's predictive accuracy.

$$SS_{\text{total}} = \sum_{i=1}^{n} (y_i - \bar{y})^2 \tag{1}$$

$$SS_{\text{residual}} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{2}$$

$$R^2 = 1 - \frac{SS_{\text{residual}}}{SS_{\text{total}}} \tag{3}$$

$$\text{R-squared} = 0.700563202695799 \tag{4}$$

Where:

- y_i represents the actual values,
- $\bar{y}$ is the mean of the actual values,
- $\hat{y}_i$ are the predicted values,
- SS_{total} is the total sum of squares,

- SS_{residual} is the sum of squared residuals.

We could see that the R squared value equal to the 0.700563202695799 suggesting that the model captures a substantial portion of the variability in the data and demonstrates that the model performs well in explaining the variability of the pct in the test data, reflecting strong predictive capabilities.

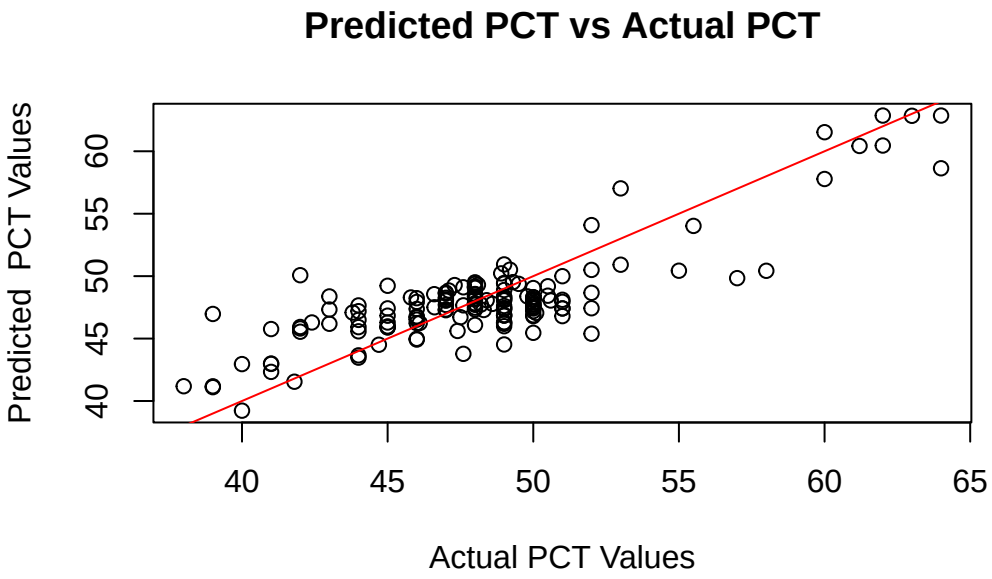

Figure 11: Predicted vs Actual Percentage Support (pct) for Kamala Harris in the Bayesian Model

By visual checking actual pct and predicted pct plot in Figure 11, the points are plotted against a 45-degree line, which represents the ideal scenario where predicted values match the actual values perfectly.

In addition, the residual plot Figure 12 are fairly centered around zero with no major trend, suggesting that the model is not heavily biased in its predictions.

In conculsion, the visual checks from the predicted vs. actual plot and the residual plot, there is no strong evidence that the Bayesian model is overfitting. It appears to generalize well to the data it was trained on without showing signs of capturing noise or irrelevant patterns.

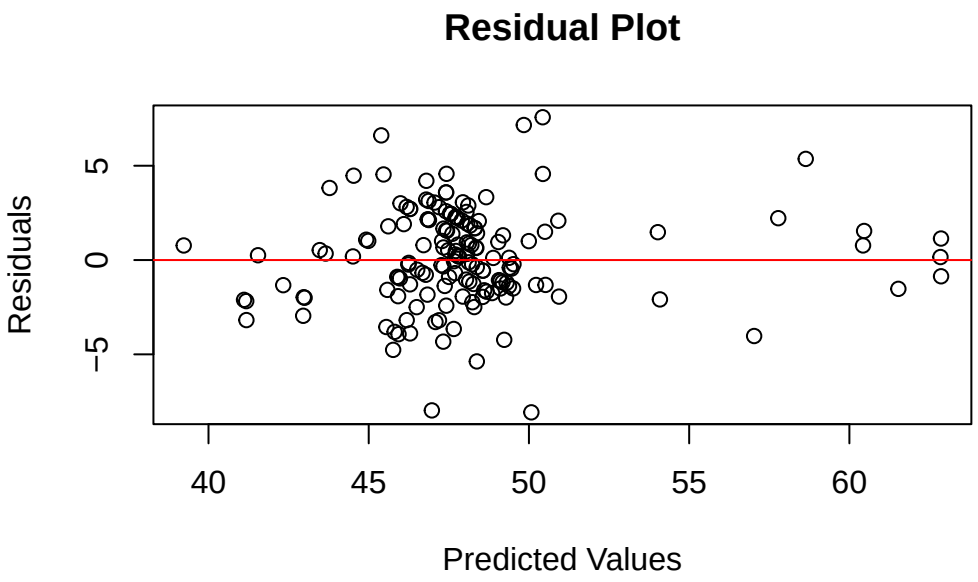

Figure 12: Residual Plot for the Bayesian Hierarchical Model Predictions

3.4.3 Model justification

To predict the support percentage (PCT) for Harris and Trump, we employed a comprehensive modeling strategy involving Simple Linear Regression (SLR), Multiple Linear Regression (MLR), and a Bayesian hierarchical model. The SLR model provided a foundational analysis of the relationship between PCT and pollscore, highlighting its limited predictive power due to a lack of complexity. The MLR model improved on this by incorporating additional predictors such as days taken from the election, sample size, methodology, and state, enhancing its predictive accuracy and capturing interactions among variables. The Bayesian hierarchical model further refined our approach, incorporating prior knowledge and accounting for variability at group levels (e.g., methodology and state) through random intercepts. This model provided robust uncertainty quantification through credible intervals and demonstrated strong predictive performance with an R-squared value around 0.70, indicating it effectively explained data variability. By leveraging the strengths of these models, particularly the Bayesian approach's interpretability and robustness, we achieved reliable predictions and a deeper understanding of the factors influencing support percentages.

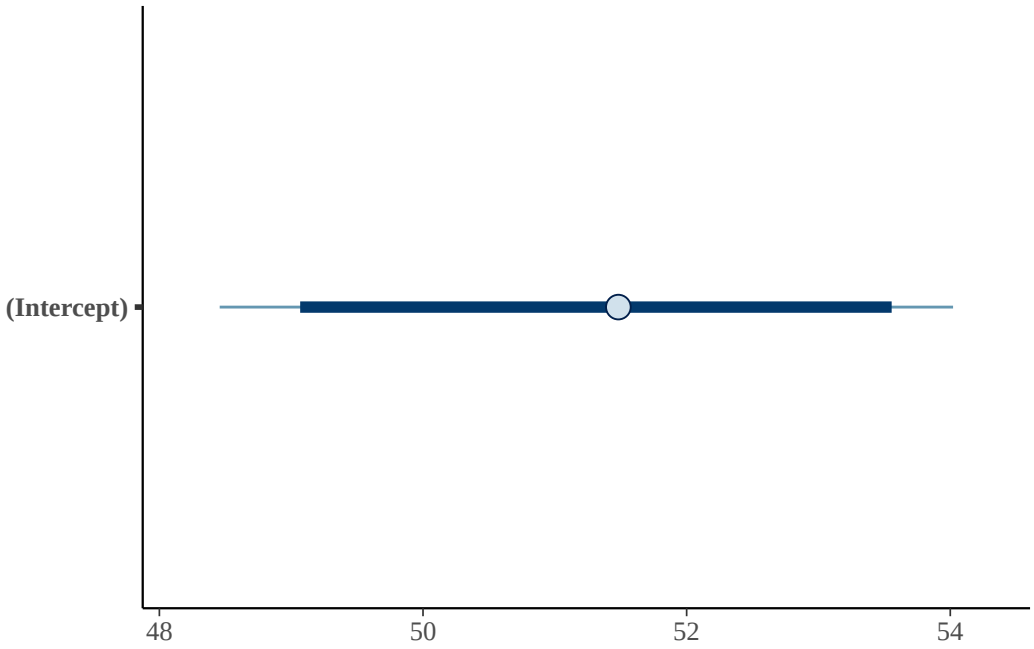

Figure 13: Estimated Posterior Distribution for the Intercept Parameter in the Bayesian Model

4 Result

Figure 13 displays the estimated posterior distribution for the intercept parameter in the Bayesian model. The intercept represents the baseline level of support, expressed as a percentage, for Kamala Harris in the U.S. election data used in the analysis. The plot features a point estimate (depicted by the central dot) that indicates the mean or median of the posterior distribution for the intercept, and a horizontal line showing the 95% credible interval, which signifies the range within which the true value of the intercept is likely to fall with 95% probability.

The estimated intercept, which represents the baseline percentage of support for Kamala Harris in the U.S. election data analyzed, is centered around 52%, with a credible interval spanning approximately from 49.5% to 53.5%. The relatively narrow width of the interval implies a certain degree of confidence in the estimate, indicating that the data used in the model provided a clear signal for the intercept's value.

Table 2: Bayesian Model Result Summary

Actual_PCT	Predicted_PCT	Lower_CI	Upper_CI
47.6	49.12357	44.41537	53.89101
48.1	49.17340	44.44684	53.88144

Table 2: Bayesian Model Result Summary

Actual_PCT	Predicted_PCT	Lower_CI	Upper_CI
48.6	47.81670	43.05068	52.51569
49.3	47.69269	42.96014	52.45805
48.1	48.04330	43.46325	52.57081
48.4	48.02722	43.07835	52.72522
46.8	48.87911	44.01774	53.56033
47.3	48.87684	44.10867	53.55697
48.1	49.14695	44.26011	53.87345
48.4	49.04718	44.33357	53.86598

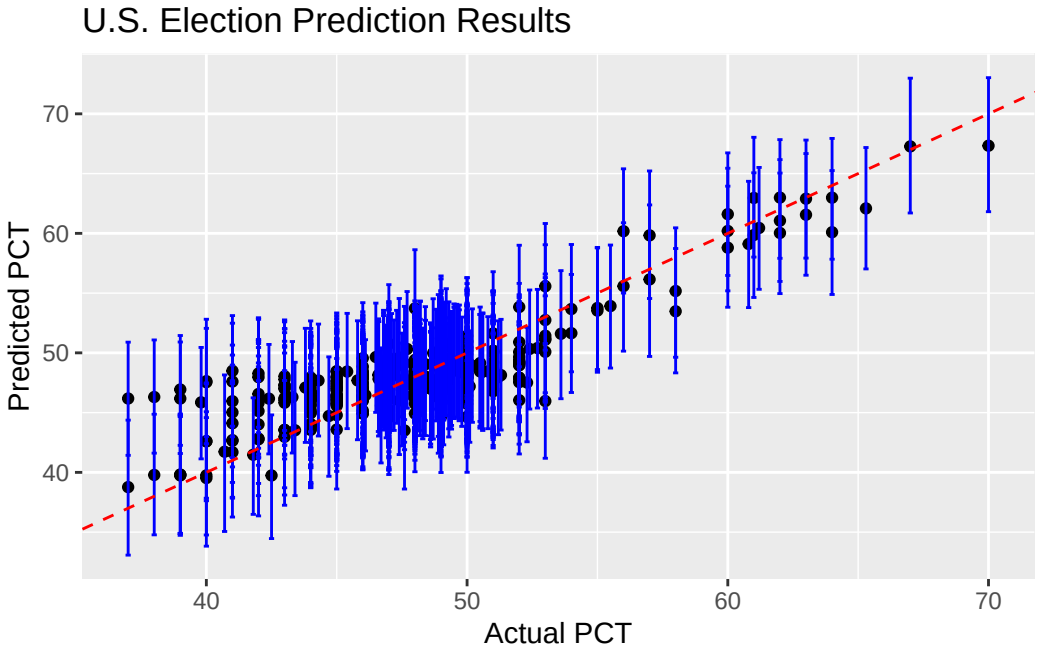

Figure 14: U.S. Election Prediction Results

Figure 14 shows that he model's predictive performance for Harris's election data appears generally robust, demonstrating a strong alignment between actual and predicted vote percentages. The majority of the data points cluster around the line, indicating reliable predictions, some variance is evident, particularly in the lower and higher actual percentage ranges. This suggests that while the model performs well on average, it may have larger uncertainty in its predictions at the extremes. Table 2 also highlights that most predicted values align closely with actual values, supporting the model's robustness in capturing the overall trend of election outcomes.

5 Discussion

5.1 Summary of Findings

In this paper, we developed a hybrid election forecasting model to predict the 2024 U.S. presidential election results. Our model provides a stable and accurate forecast compared to traditional methods by combining poll-of-polls aggregation with additional factors such as poll score, sample size, days from the election, state, and methodology.

Our approach in models applies simple linear regression, multiple linear regression and Bayesian hierarchical modeling that includes both fixed and random effects, capturing variations in poll reliability. The results shows that a model accounting for poll quality, timing, and methodology can produce forecasts that are more resilient to biases found in individual polls.

5.2 Insights on Polling and Election Dynamics

One of the main takeaways from this study is the impact of poll quality on election forecasts. We weight and filtered the polls by their reliability. This approach highlights the importance of screening and adjusting for poll reliability in election forecasts, particularly in a high-stakes environment where even minor polling biases can influence public perception and campaign strategies.

5.3 Limitations of the Model

While this model presents a more comprehensive approach, there are limitations. First, Any systematic biases inherent in those polls can carry over into the model's predictions because we are using existing poll data. For instance, if certain demographic groups are consistently underrepresented, our model may not fully capture their impact on election outcomes.

Second, the model only reflect current voting dynamics as it is based on historical data. This model does not account for short-term variability such as Biden's exit. This limitation is important in the final weeks before an election when small shifts in polling data can produce exaggerated effects.

On top of that, the results can be sensitive to the choice of priors. We choose normal(0, 2.5) as we believe it could stabilize the estimates by being informative enough to guide the posterior distribution without overpowering the data. However, if the chosen prior does not align well with the true underlying distribution, it may lead to biased results or overly wide credible intervals.

Also, as it is a model based on a dataset, it is challenging to capture all the real-life features and dynamics. The model might not include all relevant predictors or interactions due to

limitations in data availability or complexity, which can lead to incomplete representations of the factors influencing voting behavior. Additionally, without strong regularization techniques, the model may become prone to overfitting, particularly when using complex hierarchical structures or including numerous predictors. This overfitting can reduce the model's generalizability to new or unseen data.

Moreover, errors and offsets inherent in polling data, such as response bias, nonresponse adjustments, and sampling variability, can propagate into the model's results. These aspects introduce an additional layer of uncertainty that can affect the model's reliability and predictive performance. While Bayesian methods provide a robust framework to incorporate uncertainty, the final outputs must be interpreted cautiously, acknowledging these underlying limitations.

5.4 Future Directions

To enhance the reliability and robustness of this forecasting model, future work should focus on validating its performance across multiple election cycles. We includes previous year's polling dataset in this R project. By applying the model to past elections, researchers can assess its accuracy in different political contexts and electoral dynamics. This approach would show any limitations specific to certain election conditions, such as shifts in voter demographics or the influence of emerging media platforms on public opinion. Expanding the model's validation across multiple election cycles could help verify its robustness and adaptability, ultimately refining its accuracy and reliability for future applications in election forecasting.

Appendix

A FiveThirtyEight Licenses

FiveThirtyEight's data sets are used and modified by us under the Creative Commons Attribution 4.0 International License.

B Trump Voter Prediction Model

The multiple linear regression model (MLR) for Donald Trump will use the same variables, formula, and Bayesian approach as the one for Harris. Likewise, the Trump dataset is also split into training and testing data. The model outputs are below.

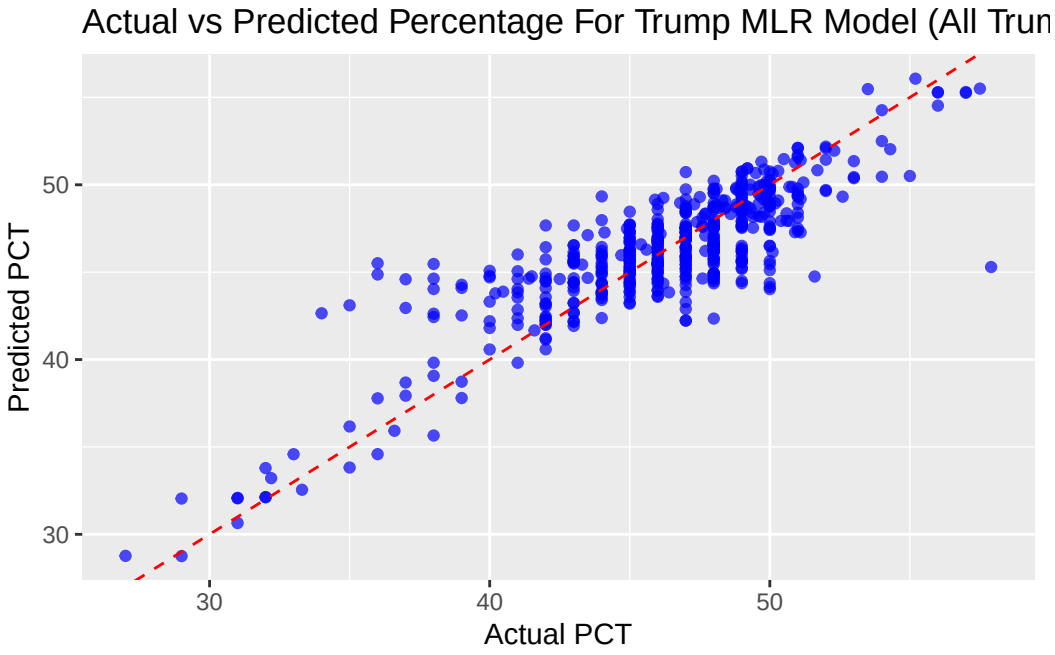

Figure 15: MLR Trump Model Accounts For A Large Amount of Variability in Voter Percentage

From Figure 15 it is clear that the model accounts for large amount of the variance in Trump's voter percentage as the data points appear close to the prediction (red line). Furthermore, the distance between the prediction and the actual values do not appear to follow a pattern, suggesting that the error is due to randomness and not model bias.

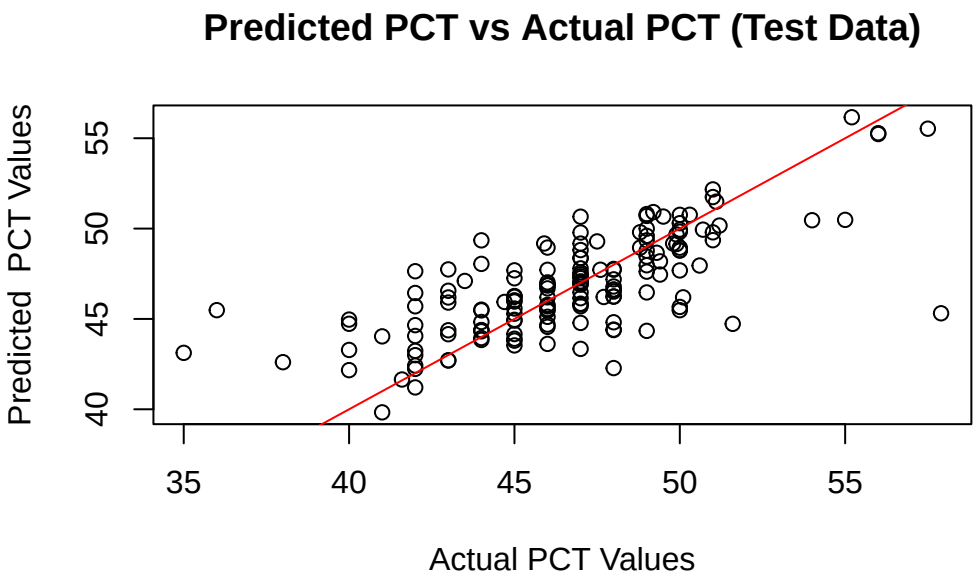

Figure 16: MLR Trump Model does not Appear to Overfit

Based on Figure 16, the test data predictions are also close to the actual values. This suggests that model can generalization to outside data. Similarly, the distance between the prediction and the actual values do not appear to follow a pattern, suggesting that the error is due to randomness and not model bias.

Based on Figure 17, the model expects Trump to win slight less than 45% of the popular vote and the 95% confidence interval ranges from around 42.5% to 47.2%. This confidence interval is slightly larger than the model for Harris, implying that the Trump polling data might be less reliable.

C Election Prediction

Our prediction process consists of two primary components. First, we develop models for both Trump and Harris based on the variables outlined in Section 3. This involves partitioning the dataset into training and testing subsets. Next, we further divide the testing dataset into swing states and other competitive races. We then input this test data into the respective models to generate predictions. By averaging these predictions, we can calculate the expected voter percentage for each candidate in each state. The candidate with the higher percentage is deemed the winner for that state.

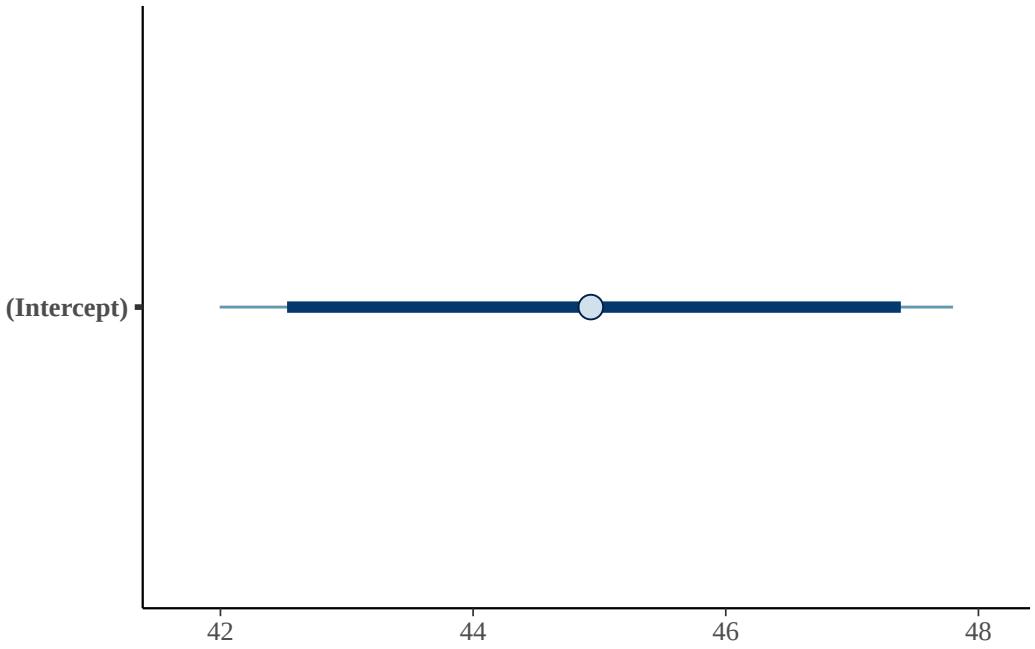

Figure 17: Donald Trump Expected to Recieve Approximatley 45% of the Vote

We generated predictions for the following states: Arizona, Nevada, Georgia, Pennsylvania, Michigan, Minnesota, Wisconsin, Florida, Texas, Maine CD-2, Nebraska CD-2, New Hampshire, Ohio, Virginia, North Carolina, and Iowa. Winners for other states were determined based on historical trends and predictions from sources like CNN (n.d.). Most states without predictions are strongly Republican or Democratic, so their absence is not expected to significantly impact prediction validity.

Table 3: Kamala Harris Wins Most of the Swing States

State	Harris Predicted Percentage	Trump Predicted Percentage	State Winner
Arizona	46.61971	49.18955	Trump
Florida	42.85817	50.73444	Trump
Georgia	47.22363	48.87560	Trump
Iowa	48.42492	43.34445	Harris
Maine CD-2	47.28499	49.27631	Trump
Michigan	47.57326	47.00349	Harris
Minnesota	48.62863	43.65465	Harris
Nebraska CD-2	49.96985	42.08998	Harris
Nevada	49.33669	47.10325	Harris

Table 3: Kamala Harris Wins Most of the Swing States

State	Harris Predicted Percentage	Trump Predicted Percentage	State Winner
New Hampshire	50.76318	42.66232	Harris
North Carolina	48.67539	47.74355	Harris
Ohio	43.97371	51.01273	Trump
Pennsylvania	48.19875	47.17264	Harris
Texas	44.96875	50.27858	Trump
Virginia	49.19355	43.20916	Harris
Wisconsin	48.45160	46.44955	Harris

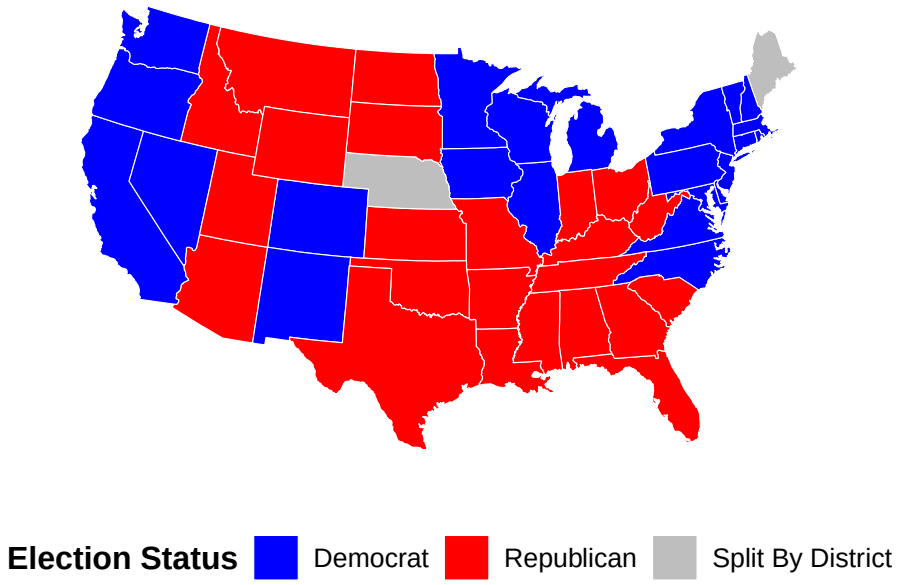

Figure 18: Kamala Harris is Predicted to be the 47th President of the United States

According to Figure Figure 19, Kamala Harris is predicted to be the 47th President of the United States. There are also a few states with predictions not visible in the map, we will describe those predicts them below.

In Maine, Harris is projected to win the state's overall delegates and District 1, while District 2 is expected to go to Trump. In Nebraska, Trump is expected to win the state's delegates along with Districts 1 and 3, while Harris is predicted to win District 2. Additionally, Trump is projected to win Alaska, and Harris is expected to win Hawaii.

Overall, the predictions indicate that Harris will receive 298 delegates, while Trump will receive 240 delegates.

D Overview of American Electon System

D.1 Brief Description of American Federal Goverment

The American federal or national government is split into three branches, executive., legislative, and judical The executive branch includes the president and the military. The legislative branch include two subgroups, the House of Representatives and the Senate. These two subgroups create laws. Every state has two Senators and one Representative per approximately 750,000 people. The judicial branch is court system.

D.2 What is the Electoral College

The Electoral College is the system used in the United States to elect the president and vice president. Instead of a direct popular vote, each state is allocated a certain number of electors based on its representation in Congress (the total of its Senators and Representatives). When voters cast their ballots, they are actually voting for a slate of electors pledged to a candidate. The candidate who receives a majority of electoral votes (270 out of 538) wins the presidency. This system means that winning the popular vote in a state generally results in winning all of that state's electoral votes. The only exception are the states of Maine and Nebraska, who award electoral votes by congressional district, with two additional votes given to the statewide winner.

The Electoral College results in some states being unnecessary to campaign in, as their strong historical voting pasterns towards either Democrats or Republicans make them unlikely to change, regardless of campaign efforts. Therefore, for statisticians, polling information from these states may not be that useful when trying to predict the outcome of an election. On the other hand, states that can vote either Democratic or Republican (swing states) are immensely important when predicting an election. As a result, campaigns spend hundreds of millions of dollars campaigning and understanding voters there.

The state statuses presented in this map are based on evaluations from CNN (n.d.), Fox News (2024), and NBC News (n.d.), which are generally agreed upon within the American political community. These organizations assessed historical voting patterns and recent polling data to derive their conclusions.

Notably, Nebraska and Maine are indicated in gray due to their delegates being split by district. In Nebraska, the state overall is projected to lean Republican, with the first and third districts also strongly favoring Republican candidates, while the second district leans Democratic. In Maine, the overall expectation is a Democratic leaning, consistent with its

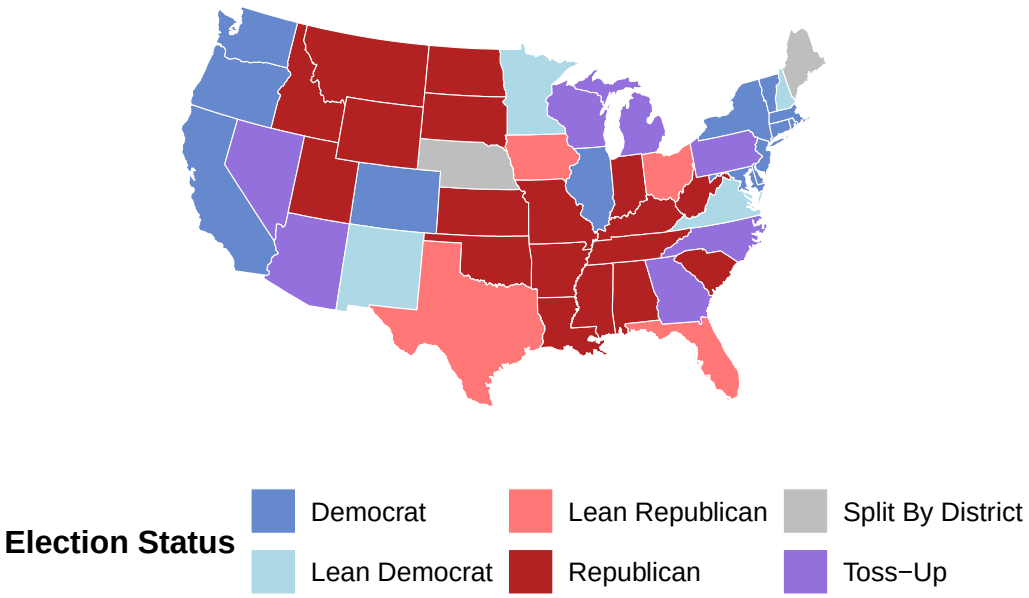

Figure 19: 2024 U.S. Presidential Election State Forecast Map

first district, though the second district leans Republican. Furthermore, Alaska and Hawaii are not in the map. Alaska is strongly favoring Republicans while Hawaii strongly favors democrats.

The seven generally agreed upon swing states are Pennsylvania, North Carolina, Georgia, Arizona, Nevada, Wisconsin, and Michigan with Texas, Florida, Nebraska District Two, Maine District 2, and Minnesota as the next closest races.

E States Poll Count

Table 4: Polls included in Analysis Per State

State	Number of Polls In Analysis
Alabama	0
Alaska	0
Arizona	17
Arkansas	0
California	4
Colorado	0

Table 4: Polls included in Analysis Per State

State	Number of Polls In Analysis
Connecticut	1
Delaware	0
Florida	6
Georgia	17
Hawaii	0
Idaho	0
Illinois	0
Indiana	1
Iowa	1
Kansas	0
Kentucky	0
Louisiana	0
Maine	2
Maryland	2
Massachusetts	3
Michigan	17
Minnesota	5
Mississippi	0
Missouri	2
Montana	2
Nebraska	2
Nevada	8
New Hampshire	5
New Jersey	0
New Mexico	2
New York	3
North Carolina	23
North Dakota	0
Ohio	4
Oklahoma	0
Oregon	0
Pennsylvania	27
Rhode Island	2
South Carolina	1
South Dakota	0
Tennessee	0
Texas	6
Utah	0
Vermont	1

Table 4: Polls included in Analysis Per State

State	Number of Polls In Analysis
Virginia	4
Washington	1
West Virginia	0
Wisconsin	19
Wyoming	0
National	75

F The Ideal Survey

F.1 Objective

The research team has developed a survey and distribution methodology with a hypothetical budget of $100,000. This objective is to create a polling system that accurately predicts the 2024 United States Election.

F.2 Sampling Frame

Given the monetary constraint, the team's first decision is to mostly include swing states and districts. These are Nevada, Arizona, Wisconsin, Michigan, Pennsylvania, North Carolina, and Georgia. Swing states are the primary focus of the poll because they disproportionately affect the outcome of the election based on the explanation in Section D.2. However, the poll will also include Texas, Florida, and Minnesota, Nebraska District 2, and Maine District 2, as they have the potential to one by either party but are not as likely to change as the previously mentioned swing states. Nevertheless, it is important to note that to save costs, the team will focus on polling true swing states compared to Texas, Florida and Minnesota.

F.3 Survey Sending Process

The team will then find and use multiple databases, such as USPS (n.d.), to find addresses, telephones, and emails of potential voters in those states. According to Pew Research Center (n.d.), gathering this information reduces non-response and selection bias, as the team can contact the same individual through multiple mediums. Moreover, certain demographics may prefer specific communication forms; for instance, the elderly may prefer phone or paper mail polls over text or email. Additionally, to encourage individuals to complete the survey, one in every 50 participants will have a chance at winning $20.

F.4 Sampling Methodology

Each state will have its own poll, but the polling methodology and the survey given will remain the same. Specifically, the polls will use stratified random sampling to make sure that participants reflect each state's counties in proportion to their population sizes. Ideally, the stratification would also ensure race, gender, age, and other socioeconomic factors are also accounted for. However, these factors are almost impossible to determine while sending the survey. Therefore, the research team decided to ask demographic questions directly inside the survey. After data collection, the team will weigh data based on demographics. For example, if a certain county has a 30% black population (this statistic will be determined from the US census), but only 15% of the survey participants are black, then the pollsters may decide to count each black participant's responses twice.

Furthermore, the team will aim to sample approximately 1,000 people from each true swing state. This sample size is large enough to provide reliable conclusions but not so large that it resembles sampling with replacement. For true swing states, the team aims to survey 1,000 individuals. However, if the final sample size falls short of this target, it is not considered a significant issue.

F.5 Survey Implementation & Question Creation

The team's ideal survey will be made using Qualtrics (n.d.). It will attempt to ensure four things: no leading questions, no question order bias, no answer order bias, and the survey should identify non-engaged participants. To identify participants who are not engaged, the research team has decided to add a worthless question. This question is extremely simple and clearly has one correct answer. Therefore, if a participant selects the wrong answer, they most likely are not engaged with the survey and their responses should be removed from the final data. An example of this is question number 7.

Order bias occurs when the sequence of questions subconsciously influences participants' responses. To prevent this, many questions should be randomized upon entry to the survey. However, some questions must follow a specific order, such as question 11, while others like questions 1-4 can be randomized. The team's hypothetical survey has not implemented this feature, though it is available with the Qualtrics (n.d.) paid plan.

Answer bias, like order bias, occurs when the order of answer choices influences participants' selections, with the first option often chosen more frequently. To prevent this, answer choices should be randomized upon survey entry. Questions 8 & 9 could benefit from this. Likewise, the team's hypothetical survey has not implemented this feature, but it is available in the Qualtrics (n.d.) paid plan.

A leading question occurs when a question is written in a way that suggests the user to give a certain answer. For example, "given that children are the future of our country, should we

invest more money in their education". To prevent this, the researchers have ensured questions are written in a style where no unnecessary details or opinions are added.

F.6 Survey Questions

Click this link for the Qualtrics (n.d.) survey.

List of all of the questions:

1. Select your race(s) (racial options where chosen based on Orvis (2024))
    - White
    - Black or African American
    - American Indian or Alaska Native
    - Native Hawaiian or Pacific Islander
    - Middle Eastern or North African
    - Asian
    - Other
    - Prefer not to answer
2. Please select your gender
    - Male
    - Female
    - Non-binary
    - Prefer not to say
3. Please enter your age (in numbers)

- This is a text input field. Please note that this field has the auto-validation feature set to numbers in Qualtrics (n.d.). As a result, participants can only input numbers in this field and are alert if they have not.

4. Please select the highest degree of education you have obtained
    - GED Certificate
    - High School Diploma
    - Undergraduate Degree
    - Graduate Degree (Masters/Phd)
    - None
    - Other
    - Prefer not to answer
5. Are you a registered voter for the 2024 United States Presidential Election?
    - Yes
    - No

6. Place yourself on the political spectrum
    - Far Left
    - Center Left
    - Center
    - Center Right
    - Far Right
7. Can pigs fly?
    - Yes
    - Maybe
    - No
    - Prefer Not To Say
8. What political party have you registered with?
    - Republicans
    - Democrats
    - Green Party
    - Libertarian
    - Other
    - Independent (unregistered)
9. Who will you vote for in the 2024 presidential election?
    - Democrat - Kamala Harris
    - Republican - Donald Trump
    - Green Party - Gill Stein
    - Libertarian - Chase Oliver
    - Other
    - Will not vote

F.7 Potential Problems with the Methodology And Polls

While the team's methodology and survey creates a robust system, there are potential issues. The reliance of weighting results based on a candidate's demographics can lead to error propagation. For instance, if a certain racial demographic population is only captured limitedly and that limited sample is far from representative, a few individuals in the population can have a large impact on the polls prediction of what candidate will win the state. There is also a selection bias in terms of the monetary reward. Potentially, people who like monetary rewards could be more likely to engage in the survey and could therefore exhibit certain voting or demographic characteristics that create an unrepresentative sample.

G Methodology of YouGov

YouGov's methodology documentations are seperated in two articles. The article by Bailey and Rivers (2024) documents the methodology of the 2024 election projection, while the webpage on YouGov (n.d.a) documents the general methodology of YouGov's prediction.

G.1 Population, Frame, and Sample

As Bailey and Rivers (2024) stated, the population covered by YouGov's MRP model is everyone in the national voter file, whether or not they belong to YouGov's panel. The national voter files are digital database built by commercial organizations with public government records of voters, as explained by DeSilver (2018). Voter files indicates whether someone voted in a given election, thus YouGov's population covers all voters in previous US elections.

YouGov's sampling frame consists of its online panel members. These members are part of the SAY24 project, a collaboration between Stanford, Arizona State, and Yale Universities, as stated by Bailey and Rivers (2024). YouGov collect information on respondents when they join their panel before they are invited to participate in the survey.

YouGov select the sample from the sampling frame based on their ability to match characteristics of the population of interest. YouGov interviews nearly 100,000 people in the first set of estimates. For the second set of estimates, YouGov didn't just start over with a new sample. They took the initial data from August and September and updated it with responses from more than 20,000 additional registered voters who were re-interviewed in late September and early October.

G.2 Sample Recruitment

Panelists are recruited through various online channels, including advertisements and partnerships with websites (YouGov n.d.a). They must provide demographic details upon joining, which helps in selecting representative samples for each survey. When respondents complete a survey, they are awarded points that can be exchanged for money.

G.3 Sampling Approach and Trade-offs

YouGov uses non-probability sampling due to the compensation, an approach where not every individual has an equal chance of selection (YouGov n.d.a). This method allows quick and cost-effective data collection. However, as YouGov (n.d.a) writes the panelists must have an internet connection to participate. YouGov state that there is 95% of us population with internet access, thus the sample may be less representative of certain hard-to-reach populations, such as individuals with very slow internet access or without internet access.

G.4 Non-response Handling

YouGov apply statistical weighting to adjust for the differences between the sample and target population. The weight is based on demographic characteristics such as age, gender, race and presidential vote (YouGov n.d.a). Additionally, quality control measures exclude unreliable responses to improve data accuracy. The respondents are offered a small incentive to decrease the non-response and increase participation.

G.5 Strengths and Weaknesses of the Questionnaire

YouGov's surveys are conducted online, which is very efficient for the respondents, and responses are weighted to enhance representativeness. The pollster can recruit a large amount of panelists because of the online format. Combining with online tracking technologies, the metadata provided by ther panelists can be verified easily.

As a non-probability sample, it might miss certain demographic groups not covered by the online population. While weighting improves accuracy, it cannot fully substitute the randomization found in probability sampling. Additionaly, the categories in the surbey is oversimplified with bias. For instance, in the poll result published by YouGov (n.d.b), gender is divided into Male and Female. Race is divided into White, Black, Hispanic and Other. This indicates a lack of representation.

References

Auguie, Baptiste. 2017. *gridExtra: Miscellaneous Functions for "Grid" Graphics*. https://CRAN.R-project.org/package=gridExtra.

Bailey, Delia, and Douglas Rivers. 2024. "How YouGov's MRP Model Works for the 2024 U.S. Presidential and Congressional Elections." *YouGov*. https://today.yougov.com/politics/articles/50587-how-yougov-mrp-model-works-2024-presidential-congressional-elections-polling-methodology.

Bates, Douglas, Martin Maechler, and Mikael Jagan. 2024. *Matrix: Sparse and Dense Matrix Classes and Methods*. https://CRAN.R-project.org/package=Matrix.

Blumenthal, Mark. 2014. "Polls, Forecasts, and Aggregators." *PS: Political Science and Politics* 47 (2): 297–300. http://www.jstor.org/stable/43284537.

Brilleman, SL, MJ Crowther, M Moreno-Betancur, J Buros Novik, and R Wolfe. 2018. "Joint Longitudinal and Time-to-Event Models via Stan." https://github.com/stan-dev/stancon_talks/.

CNN. n.d. "Electoral College Map 2024: Road to 270." *CNN Politics*. Accessed November 2, 2024. https://www.cnn.com/election/2024/electoral-college-map.

DeSilver, Drew. 2018. "Q&A: The Growing Use of 'Voter Files' in Studying the U.S. Electorate." *Pew Research Center*. https://www.pewresearch.org/short-reads/2018/02/15/voter-files-study-qa/.

FiveThirtyEight. 2024. "Our Data." *FiveThirtyEight*. https://data.fivethirtyeight.com.

Fox News. 2024. "2024 Presidential Power Rankings." *270toWin*. https://www.270towin.com/maps/fox-news-2024-presidential-power-rankings.

Grolemund, Garrett, and Hadley Wickham. 2011. "Dates and Times Made Easy with lubridate." *Journal of Statistical Software* 40 (3): 1–25. https://www.jstatsoft.org/v40/i03/.

Morris, G. Elliott. 2024. "Trump Leads in Swing-State Polls and Is Tied with Biden Nationally." *ABC News*. https://abcnews.go.com/538/trump-leads-swing-state-polls-tied-biden-nationally/story?id=109506070.

Müller, Kirill. 2020. *Here: A Simpler Way to Find Your Files*. https://CRAN.R-project.org/package=here.

NBC News. n.d. "Choose a Path to the White House with Our 'Road to 270' Map." *Nbcnews.com*. Accessed November 3, 2024. https://www.nbcnews.com/specials/road-to-270-electoral-college-interactive-map-2024-election/.

Orvis, Karin. 2024. "OMB Publishes Revisions to Statistical Policy Directive No. 15: Standards for Maintaining, Collecting, and Presenting Federal Data on Race and Ethnicity | OMB." *The White House*. https://www.whitehouse.gov/omb/briefing-room/2024/03/28/omb-publishes-revisions-to-statistical-policy-directive-no-15-standards-for-maintaining-collecting-and-presenting-federal-data-on-race-and-ethnicity/.

Pasek, Josh. 2015. "The Polls–Review: Predicting Elections: Considering Tools to Pool the Polls." *The Public Opinion Quarterly* 79 (2): 594–619. http://www.jstor.org/stable/24546379.

Pew Research Center. n.d. "U.S. Surveys." *Pew Research Center*. Accessed November 2, 2024.

https://www.pewresearch.org/u-s-surveys/.

Qualtrics. n.d. "Qualtrics XM - Experience Management Software." *Qualtrics*. Accessed November 3, 2024. https://www.qualtrics.com/.

R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Radcliffe, Mary, and G. Elliott Morris. 2023. "538's Polls Policy and FAQs." *ABC News*. https://abcnews.go.com/538/538s-polls-policy-faqs/story?id=104489193.

Richardson, Neal, Ian Cook, Nic Crane, Dewey Dunnington, Romain François, Jonathan Keane, Dragoș Moldovan-Grünfeld, Jeroen Ooms, Jacob Wujciak-Jens, and Apache Arrow. 2024. *Arrow: Integration to 'Apache' 'Arrow'*. https://CRAN.R-project.org/package=arrow.

Silver, Nate. 2008. "Polls Now Weighted by Sample Size." *FiveThirtyEight*. https://fivethirtyeight.com/features/polls-now-weighted-by-sample-size/.

Stan Development Team. 2024. "RStan: The R Interface to Stan." https://mc-stan.org/.

USPS. n.d. "Customer Relations and Domestic Mail." Accessed November 3, 2024. https://about.usps.com/postal-bulletin/2001/html/pb22057/a-d.html.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Wickham, Hadley, Romain François, Lionel Henry, Kirill Müller, and Davis Vaughan. 2023. *Dplyr: A Grammar of Data Manipulation*. https://CRAN.R-project.org/package=dplyr.

Xie, Yihui. 2014. "Knitr: A Comprehensive Tool for Reproducible Research in R." In *Implementing Reproducible Computational Research*, edited by Victoria Stodden, Friedrich Leisch, and Roger D. Peng. Chapman; Hall/CRC.

YouGov. n.d.a. "Methodology." Accessed October 31, 2024. https://today.yougov.com/about/panel-methodology.

———. n.d.b. "SAY24 Poll." Accessed November 3, 2024.